Perspectives on Real-time Computation of Movement Coarticulation

Frédéric Bevilacqua, Baptiste Caramiaux, Jules Françoise

Perspectives on Real-time Computation of Movement Coarticulation

Frédéric Bevilacqua
Ircam - Centre Pompidou
STMS IRCAM-CNRS-UPMC
Paris, France
frederic.bevilacqua@ircam.fr

Baptiste Caramiaux
McGill University
Montreal, QC, Canada
STMS IRCAM-CNRS-UPMC
Paris, France
baptiste.caramiaux@ircam.fr

Jules Françoise
School of Interactive Arts and Technologies
Simon Fraser University
Surrey, Canada
jfrancoi@sfu.ca

ABSTRACT

We discuss the notion of movement coarticulation, which has been studied in several fields such as motor control, music performance and animation. In gesture recognition, movement coarticulation is generally viewed as a transition between "gestures" that can be problematic. We propose here to account for movement coarticulation as an informative element of skilled practice and propose to explore computational modeling of coarticulation. We show that established probabilistic models need to be extended to accurately take into account movement coarticulation, and we propose research questions towards such a goal.

Author Keywords

Coarticulation; Movement; Gesture; Recognition; Motor Primitives

ACM Classification Keywords

H.5. Information Interfaces and Presentation (e.g. HCI): Multimedia Information Systems; G.3Probability And Statistics: Time series analysis; J.5 Arts and Humanities: Performing arts (e.g., dance, music)

INTRODUCTION

Coarticulation is a well known phenomena in speech production occurring when a sound segment is influenced by its context, such as the preceding and following sound segments in a word or sentence[1]. This problem has been widely studied and modeled for both speech recognition and generation [14]. While coarticulation has also been described in movement sequences, it remains largely overlooked. In Human-Computer Interaction (HCI), and particularly in gesture-based interaction, the phenomenon of coarticulation is often considered as a "problem", since it perturbs the performance of gestural vocabularies and can reduce recognition rate [1].

Coarticulation relies on the existence and formalization of constitutive *segments*. For instance speech is often examined as a finite number of phonological sound segments that are ordered, and which ordering is linguistic-dependent. Coarticulation can be observed and measured because the alteration of phonemes remains consistent over a large vocabulary. Considering movement, such segments become highly variable across individuals and context-dependent, making their formalization inherently more complex. Motor theorists proposed the notion of movement *primitives* [3] as basic units (typically, patterns of movement kinematics) that can be sequenced to execute a complex movement. Within this framework, movement coarticulation has been linked to motor skills that involve the selection of movement primitives, their ordering and their accurate execution [29].

In this context, an important challenge is the computational modeling of coarticulating movements with the aim of leveraging on user's motor skills rather than discarding them for the sake of accuracy in gesture recognition systems. This paper aims at discussing problems and prospects with computational modeling of coarticulation for the field of movement and computing. In particular, we propose to include aspects from motor control theories. We first recall some important references for movement coarticulation that span over different disciplines. Second we present computational approaches in interactive systems. We then illustrate typical coarticulation phenomena occurring with simple gestural inputs and the analysis of these phenomena through the lens of existing computational models. Finally we shortly discussed the results and propose a perspective for computational movement models involving coarticulation.

COARTICULATION ACROSS DIFFERENT FIELDS

Coarticulation has been studied and formalized in speech perception [30] and recognition [25, 4], communication, animation, embodied conversational agents [21] as well as sign language (both in pure synthesis and motion retrieval and data-driven sign-language synthesis) [11, 26, 17].

In motor control, coarticulation has been studied considering simple tasks and movements [15, 27, 29, 23]. Typically, coarticulation occurs in sensori-motor learning when the movement primitives (understood as basic units such as patterns of

[1]We note that the word coarticulation is also used to describe the occurrence of two different modalities, for example voice and gesture. In this paper, we used the term coarticulation as it is generally used in speech production.

movement kinematics [3]) are fused in a larger phrase, which does not appear as a series of separate events [12], but from an intermediate-level command that encompasses several events as one event, called "chunk". Chunks' boundaries are characterised by higher motor variability and probability of errors. This behavioral characterisation is supported by recent work in neuroscience proposing a hierarchical motor representation underlying expert performance (see for example the recent review of [8]). As a corollary, coarticulation is generally considered as the results of an *anticipative* behaviour: the execution of the unit is planned ahead and the movement appears to start before the end of the previous unit [29].

The concept of coarticulation has also found an echo in the study of musical gestures, and in particular instrumental gestures [13, 20, 22, 2, 12, 20]. In particular, Godøy [12] proposes an informative review of coarticulation in music performance. In music performance, the notion of small movement units can be linked to existing musical events such as musical notes or sounds. Such link between score events and segmented instrumentalists' gestures have also been examined through the use of computational model [7]. Motor theories and cognition should be considered here since instrumental gestures imply learning skilled movements and anticipation. Yet, important works remains to be conducted on this area.

Finally, in dance, motor skill learning is acknowledged as a fundamental aspect of the practice. Thus dance constitutes also a promising field for investigating coarticulation. Nevertheless, to our knowledge the notion of coarticulation in dance has been less studied from a computational perspective. Here the challenges are two-fold: to define what constitutes a movement segment that can then be used in computing systems, and to design computing systems able to understand higher-level representations of complex dance movements coherent with embodied cognitive mechanisms [16].

REAL-TIME COMPUTATION OF COARTICULATION

In this work, we consider gesture-based interactive systems in which movement analysis must be achieved in real-time. In such cases, the challenge is to identify and characterize segments from a continuous stream of motion data while simultaneously accounting for their context-dependent variations. In this section, we introduce the type of models that we are considering for real-time computation of coarticulation.

In the literature, spotting and classifying gestures in a continuous stream of movement data is usually called *continuous* gesture recognition. State-space temporal models are typically used for continuous gesture recognition because they can take into account both variability in execution and temporal dependencies in the signal [19]. For example, Conditional Random Fields (CRFs) have been shown successful for such a task [32, 18]. Moreover, considering coarticulation, CRF is able to take into account contextual information such as the preceding and following of a given segment. However, as described in Morency et al. [18], standard implementation of CRF can only be used for offline analysis on bounded continuous stream preventing for its use in interactive systems. Nevertheless, Song et al [28] extended the models by proposing an online spotting within a sliding window, but not addressing explicitly articulation effects.

Considering gesture-based interactive systems, we believe that the challenge is actually to go beyond continuous gesture recognition (i.e.spotting) in characterizing the gestures execution and in particular their coarticulation [5]. Importantly, physical movements are dynamic phenomena, which encode features directly linked to expressivity. It is particularly true for coarticulatory movements that are prone to unconscious cognition-induced changes in dynamics (e.g. chunking) as well as voluntary (conscious) continuous variations.

In our previous work, we have proposed Bayesian state-space models able to infer in real-time the gesture performed and characteristics of its execution. We have proposed two main approaches for this problem: a template-based continuous state-space model [6], and a variant of hidden Markov models [10, 9]. The former is able to track modulations of recorded templates, the latter is able to learn statistically-relevant gesture variances. As both approaches proposed two complementary views of potential variability in movement execution, we inspect in the next section how these models can inform on the coarticulatory content in gesture sequences.

A TOY EXAMPLE

This section aims to illustrate a typical case of gesture coarticulation and the associated challenges for real-time analysis and continuous gesture recognition.

Movement Measurement

We recorded a set of executions of two gestures drawn on a trackpad, using Cycling'74 Max[2] with the external *fingerpinger*[3] to measure the trajectories. We also used the library MuBu[4] for Max [24] in order to record and save the captured gesture data.

We chose to use two gestures, typically used in gesture-based interactive systems (see for example [31]). The first gesture is a "V" (Figure 1, a) the second is a "O" (Figure 1, b). These two gestures are then sequenced in two different ways: *Gesture* 1 – *Stop* – *Gesture* 2 and *Gesture* 1 – *Gesture* 2 (no pause between both gesture executions). The 4 gestures are depicted in Figure 1. Each one of these gestures is repeated 10 times.

The two different ways to perform transitions between gesture 1 and gesture 2 illustrate different aspects of coarticulation. In the first case, the coarticulatory effect is minimised since the movement stops at the transition. In the second case, coarticulatory effects intervene since the movement is not allowed to stop and fused boundaries between segments must appear.

Real-Time Inference of Coarticulated Gestures

We analyze the coarticulation between the two gestures through the two probabilistic models mentioned.

[2] `http://www.cycling74.com`

[3] by Michael and Max Egger `http://www.anyma.ch/2009/research/multitouch-external-for-maxmsp/`

[4] `http://forumnet.ircam.fr/mubu`

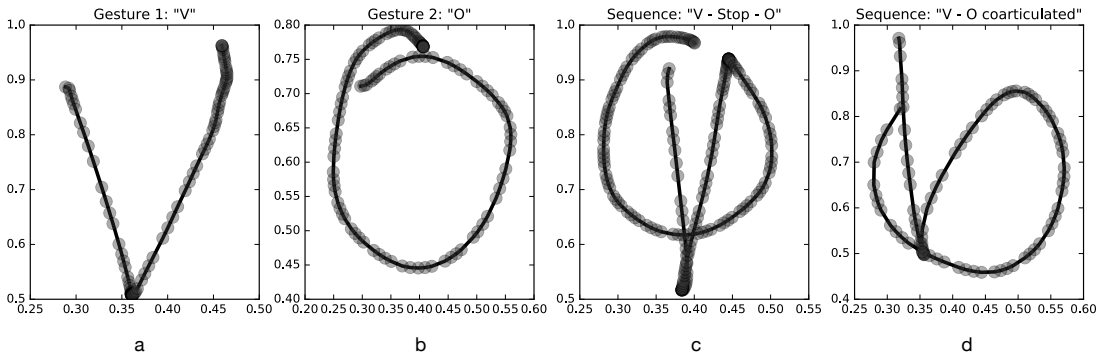


Figure 1: The gestures used in the experiment and several instances of their sequences. The two basic gestures are two-dimensional trajectories representing the symbols V (a) and O (b). Figures (c) and (d) respectively represent a sequence of gestures 1 and 2, either with a pause between, or rapidly without break.

The first model is an adaptive template-based following system [6]. In this case we learn first the whole sequence of gestures 1 and 2 without coarticulation (i.e. performed with a short stop between them) as the template (Figure 2a-left). Then, we observe how the adaptive model can match the coarticulated sequence (Figure 2a-middle). The Figure 2a also shows the alignment that is computed between each gesture segment, 1 and 2 (Figure 2a-middle and right). The adaptive following system can track the co-articulated figure, and the transition between the two gestures appears clearly on the alignment (Figure 2a-right).

Such an approach nevertheless requires to record, thus 'learn' the complete pair of gesture or at least their transition, similarly to speech where diphones are considered.

Next, we consider how the coarticulation can be modeled when a statistical model learns each gesture separately. We use a hierarchical hidden Markov model that encodes each gesture through a learning procedure [9]. Gesture 1 and 2 are learned by considering the 10 examples of the isolated performance. A 10-state HMM is built from these gestures. The analysis is performed online on the same chosen sequences.

Figure 2b shows the results of the real-time continuous gesture recognition for the coarticulated sequences.

The HMM approach is a discrete-state model which allows for a sharp transition between both gestures. Nevertheless, we observed that the transition cannot always be defined as a unique point, as it is illustrated in Figure 2b-right.

Precisely, in the case of the HMM-like approach, the model is able to consistently follow the first gesture (the time progression evolves continuously from 0 to 1). However, the transition between the two gestures results in short-term recognition errors. This problem is typical of real-time continuous gesture recognition, where the ambiguities of coarticulation results in 'jumps' of the recognition until one gesture is resolved after the transition is complete.

DISCUSSION AND PERSPECTIVES

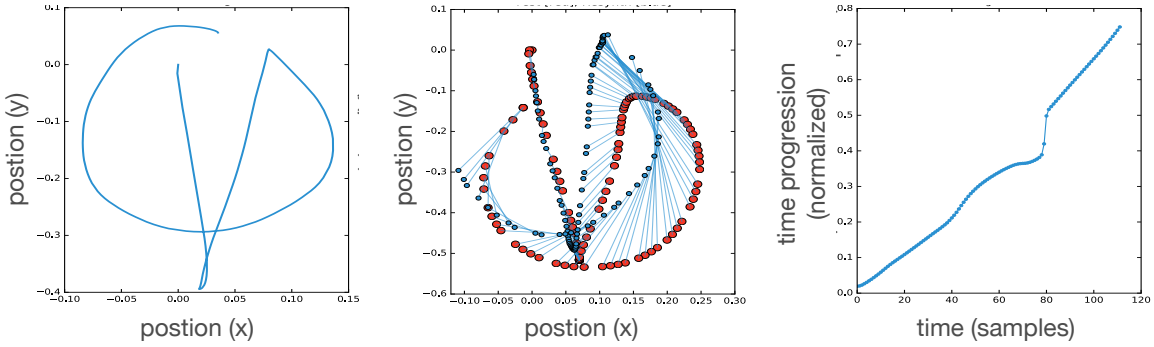
The example presented in this paper illustrated that both models can account for coarticulation in a first approximation. Importantly, boundaries are fuzzy and there is a need for an interpolation (or extrapolation) strategy. Usually, constraints on transition are imposed on the user in order to ensure better recognition results as observed in the case where a pause is respected between both gestures. Nevertheless, we advocate here to improve movement modeling to better take into account coarticulation. In the first model, interpolating would mean interpolate between two templates by allowing cross-gestural dependencies. The applications of method proposed in computer animation (e.g. Gibet et al [11]) could be interesting to evaluate in this context. In the second model, cross-gestural dependencies could also be envisaged but it would require examples of such dependencies in order to capture their structure.

It can indeed be argue that large database containing several ways of performing coarticulated gestures would improve recognition even in the presence of coarticulation. However, our point here is different: coarticulation intrinsically contains important information that we should be able to characterize *per se* in our computational models, and exploit in the interaction. In particular, coarticulation can inform on the degree of expertise in skilled movements, as usually found in music and dance. Moreover, it would be beneficial to relate computational models to the notion of chunking [8]. Coarticulation typically occurs within a chunk, and segmentation marks could be detected between chunks.

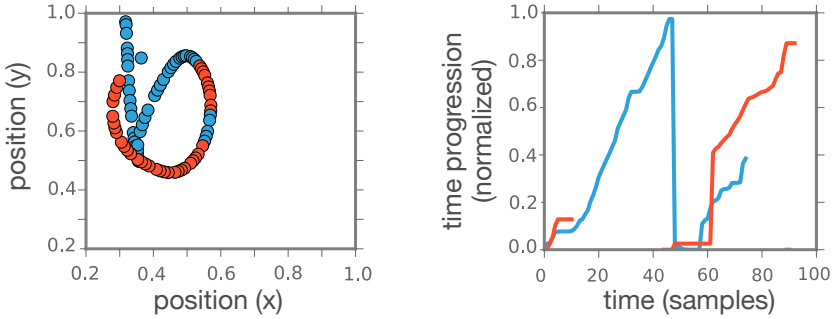
It is important to note that the example we provide in 2D is only representative of the research questions we propose. The use of 3D trajectory and other movement modalities such as acceleration data would pose additional issues. The generalisation of the methodology to a 3D movement remains an important goal of this research.

As a perspective, we propose to take into account the following points for computational movement models:

- Movement anticipation implies to take into account the influence of precedent segments for describing forthcoming segments;



(a) Tracking test with the template-based method using particle filtering. The blue figure is the template performed with a pause (*V – Stop – 0*) and the red figure shows the performed coarticulated gesture (no pause between *V* and *0*). The central figure shows the spatial alignment between the two figures and the right figure shows the temporal alignment where the transition is clearly visible.



(b) Recognition test with the Hierarchical HMM. The left figure shows the coarticulated gestures (no pause between *V* and *0*). The color corresponds to the recognition results: blue is for *V* and red for *0*. The right figure shows the recognition over time: the blue curve shows the time progression (as decoded by the model) during the *V*, followed by time progression during the *0*. In both figures, the transition is ambiguous.

Figure 2: Results of (a) gesture tracking using template-based approach and (b) continuous gesture recognition using a Hierarchical Hidden Markov Model)

- Low-level segments are concatenated in longer phrases through practice and learning, which produce movement *variations* over time altering the initial shapes of the segments and containing expressive features. Thus, movement features and vocabularies must be considered as evolving over time and prone to user idiosyncrasies.
- As fused boundaries between segments might appear through practice, segmentation should be adaptive. Hybrid segmentation-interpolation approaches should be considered.

One approach is to consider a fully Bayesian movement representation that would be take into account uncertainty cross-gestures, various time scales and time courses in anticipation. Other approaches are however possible and we hope that this 'perspective' paper could contribute to trigger important discussions on this topic.

ACKNOWLEDGMENTS

This project has received funding from the European Union?s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 659232, the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02, and by the movingstories research partnership (SSHRC Award 31639884).

REFERENCES

1. Bhuyan, M., Kumar, D. A., MacDorman, K. F., and Iwahori, Y. A novel set of features for continuous hand gesture recognition. *Journal on Multimodal User Interfaces 8*, 4 (2014), 333–343.
2. Bianco, T., Freour, V., Rasamimanana, N., Bevilaqua, F., and Caussé, R. On gestural variation and coarticulation effects in sound control. In *Lecture Notes in Computer Science*, vol. 5934. Springer, 2009, 134–145.
3. Bizzi, E., Mussa-Ivaldi, F. A., and Giszter, S. Computations underlying the execution of movement: a biological perspective. *Science 253*, 5017 (1991), 287–291.
4. Bush, B. O. *Modeling coarticulation in continuous speech*. PhD thesis, Oregon Health and Science University, 2015.

5. Caramiaux, B., Bevilacqua, F., and Tanaka, A. Beyond recognition: using gesture variation for continuous interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, ACM (2013), 2109–2118.
6. Caramiaux, B., Montecchio, N., Tanaka, A., and Bevilacqua, F. Adaptive gesture recognition with variation estimation for interactive systems. *ACM Transactions on Interactive Intelligent Systems (TiiS) 4*, 4 (2015), 18.
7. Caramiaux, B., Wanderley, M. M., and Bevilacqua, F. Segmenting and parsing instrumentalists' gestures. *Journal of New Music Research 41*, 1 (2012), 13–29.
8. Diedrichsen, J., and Kornysheva, K. Motor skill learning between selection and execution. *Trends in Cognitive Sciences 19*, 4 (2015), 227–233.
9. Françoise, J., Roby-Brami, A., Riboud, N., and Bevilacqua, F. Movement sequence analysis using hidden markov models: A case study in tai chi performance. In *Proceedings of the 2Nd International Workshop on Movement and Computing*, MOCO '15, ACM (Vancouver, British Columbia, Canada, 2015), 29–36.
10. Françoise, J., Schnell, N., Borghesi, R., and Bevilacqua, F. Probabilistic models for designing motion and sound relationships. In *Proceedings of the 2014 International Conference on New Interfaces for Musical Expression* (2014), 287–292.
11. Gibet, S., Lebourque, T., and Marteau, P.-F. High-level specification and animation of communicative gestures. *Journal of Visual Languages & Computing 12*, 6 (2001), 657–687.
12. Godøy, R. I. Understanding coarticulation in musical experience. In *Sound, Music, and Motion*. Springer, 2013, 535–547.
13. Godøy, R. I., Jensenius, A., and Nymoen, K. Chunking in music by coarticulation. *Acta Acustica united with Acustica 96*, 4 (2010), 690–700.
14. Hardcastle, W. J., and Hewlett, N. *Coarticulation: Theory, data and techniques*. Cambridge University Press, 2006.
15. Iskarous, K., Mooshammer, C., Hoole, P., Recasens, D., Shadle, C. H., Saltzman, E., and Whalen, D. The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance. *The Journal of the Acoustical Society of America 134*, 2 (2013), 1271–1282.
16. Kirsh, D. Embodied cognition and the magical future of interaction design. *ACM Transactions on Computer-Human Interaction (TOCHI) 20*, 1 (2013), 3.
17. Li, S., Wang, L., and Kong, D. Synthesis of sign language co-articulation based on key frames. *Multimedia Tools and Applications 74*, 6 (2015), 1915–1933.
18. Morency, L.-P., Quattoni, A., and Darrell, T. Latent-dynamic discriminative models for continuous gesture recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, IEEE (2007), 1–8.
19. Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
20. Palmer, C., and Deutsch, D. Music performance: Movement and coordination. *The psychology of music 3* (2013), 405–422.
21. Pelachaud, C. *Communication and coarticulation in facial animation*. PhD thesis, University of Pennsylvania, 1991.
22. Rasamimanana, N. H., and Bevilacqua, F. Effort-based analysis of bowing movements: evidence of anticipation effects. *The Journal of New Music Research 37*, 4 (2009), 339 – 351.
23. Säfström, D., Flanagan, J. R., and Johansson, R. S. Skill learning involves optimizing the linking of action phases. *Journal of neurophysiology 110*, 6 (2013), 1291–1300.
24. Schnell, N., Röbel, A., Schwarz, D., Peeters, G., and Borghesi, R. MuBu & Friends - Assembling Tools for Content Based Real-Time Interactive Audio Processing in Max/MSP. In *Proceedings of the International Computer Music Conference (ICMC)* (2009).
25. Schultz, T., and Wand, M. Modeling coarticulation in emg-based continuous speech recognition. *Speech Communication 52*, 4 (2010), 341–353.
26. Segouat, J. A study of sign language coarticulation. *SIGACCESS Access. Comput.*, 93 (Jan. 2009), 31–38.
27. Shah, A., Barto, A. G., and Fagg, A. H. A dual process account of coarticulation in motor skill acquisition. *Journal of motor behavior 45*, 6 (2013), 531–549.
28. Song, Y., Demirdjian, D., and Davis, R. Continuous body and hand gesture recognition for natural human-computer interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS) 2*, 1 (2012), 5.
29. Sosnik, R., Hauptmann, B., Karni, A., and Flash, T. When practice leads to co-articulation: the evolution of geometrically defined movement primitives. *Experimental Brain Research 156*, 4 (2004), 422–438.
30. Viswanathan, N., Magnuson, J. S., and Fowler, C. A. Information for coarticulation: Static signal properties or formant dynamics? *Journal of Experimental Psychology: Human Perception and Performance 40*, 3 (2014), 1228.
31. Wobbrock, J., Wilson, A., and Li, Y. Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, ACM (2007), 159–168.
32. Yang, R., and Sarkar, S. Detecting coarticulation in sign language using conditional random fields. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 2, IEEE (2006), 108–112.